

Discrete and Soft Prompting for Multilingual Models

Mengjie Zhao and Hinrich Schütze

CIS, LMU Munich, Germany

mzhao@cis.lmu.de

Abstract

It has been shown for English that discrete and soft prompting perform strongly in few-shot learning with pretrained language models (PLMs). In this paper, we show that discrete and soft prompting perform better than finetuning in multilingual cases: Crosslingual transfer and in-language training of multilingual natural language inference. For example, with 48 English training examples, finetuning obtains 33.74% accuracy in crosslingual transfer, barely surpassing the majority baseline (33.33%). In contrast, discrete and soft prompting outperform finetuning, achieving 36.43% and 38.79%. We also demonstrate good performance of prompting with training data in multiple languages other than English.

1 Introduction

Prompting strongly outperforms finetuning (Devlin et al., 2019) when adapting pretrained language models (PLMs; Devlin et al. (2019); Conneau et al. (2020)) to downstream tasks in the low-resource regime (Brown et al., 2020; Schick and Schütze, 2021; Gao et al., 2020; Tam et al., 2021; Le Scao and Rush, 2021), i.e., *few-shot learning*, a more realistic scenario than having tens of thousands of annotations, *even for English* (Yu et al., 2018; Yin et al., 2020; Ram et al., 2021).

In contrast to finetuning, which learns discriminative classifiers for tasks like natural language inference (NLI; Dagan et al. (2006); Bowman et al. (2015)), prompting reformulates the classification task to generative text-to-text (Raffel et al., 2020) or cloze-style (McCann et al., 2018; Brown et al., 2020) queries which are given to a PLM to answer. For example, the NLI task of assigning premise “They whinnied, eyes wide” and hypothesis “Their eyes were open wide” to class “entailment” can be reformulated as:

They whinnied, eyes wide . Question: Their eyes were open wide ? Answer: _ .

The PLM is requested to fill in, for the blank (), the word “yes”, which is mapped to “entailment”.

Prompting makes a human description of the task available in learning. Also, “filling in the blank” is well aligned with the pretraining objective (masked/autoregressive language modelling (Devlin et al., 2019; Radford et al., 2019; Yang et al., 2019)), likely to deliver better performance in few-shot learning (Ram et al., 2021).

In this paper, we investigate the effectiveness of prompting in multilingual tasks, which – despite the success of prompting in English – is largely unexplored. We address two main research questions: (RQ1) Does the strong few-shot performance of prompting transfer to other languages from English? (RQ2) As the cost of few-shot non-English annotations is affordable (Garrette and Baldridge, 2013; Lauscher et al., 2020; Zhao et al., 2021), can we directly prompt PLMs in languages other than English or do we have to go through the (generally best resourced) intermediary of English?

In this work, we systematically compare two popular prompting methods – discrete and soft prompting – with finetuning in the few-shot multilingual NLI task and show that *prompting is superior*: (i) The strong few-shot learning performance of prompting transfers to other languages from English: It outperforms finetuning in crosslingual transfer (RQ1; §5.1). (ii) Directly querying the multilingual PLM with few-shot non-English prompts achieves competitive performance, without relying on crosslingual transfer from English (RQ2; §5.2).

2 Related Work

GPT3 (Brown et al., 2020) succeeds in few-shot NLU tasks with “in-context learning”: A natural language prompt describing the NLU task is prepended to an input example; GPT3 is then capable of making accurate predictions *without updating its parameters*. However, the number of

parameters in GPT3 is prohibitively large (175B).

Integrating gradient descent into prompting, smaller (w.r.t. GPT3) PLMs also achieve good few-shot performance. Like GPT3, **discrete prompting** uses natural language to describe NLU tasks. Schick and Schütze (2021), Tam et al. (2021), Le Scao and Rush (2021) use human-designed prompts. Gao et al. (2020) leverage T5 (Raffel et al., 2020) to generate prompts. Shin et al. (2020) use extra training data to search tokens for constructing the prompts. Discrete prompting naturally inherits interpretability from the task descriptions.

Soft prompting relaxes the constraint that a prompt needs to be composed of discrete tokens. Instead, it learns the prompt in the continuous space with SGD. Qin and Eisner (2021) and Zhong et al. (2021) learn soft prompts eliciting more knowledge (Petroni et al., 2019) from PLMs than discrete prompts. Similar to soft prompting but with the PLM being frozen, Li and Liang (2021) propose prefix-tuning to encourage PLMs to solve generation tasks with high parameter-efficiency (Houlsby et al., 2019; Zhao et al., 2020). Lester et al. (2021) demonstrate that soft prompting benefits from scaling up the number of PLM parameters. Liu et al. (2021) show that GPT (Radford et al., 2019) can solve NLU tasks (Wang et al., 2019) with soft prompting.

All of this work focuses on English. We show that discrete and soft prompting perform better than finetuning in few-shot crosslingual natural language inference (XNLI; Conneau et al. (2018)) with multilingual PLMs (XLM-RoBERTa; Conneau et al. (2020)). We conduct experiments on NLI because it is one of the most representative and challenging NLU tasks (Dagan et al., 2006; Bowman et al., 2015), and has been commonly used in prior work on prompting.

3 Method

3.1 Finetuning

We follow the standard finetuning method (Devlin et al., 2019): A linear classifier layer is initialized and stacked on top of the PLM; the whole model is then trained on the few-shot NLI dataset (§4).

3.2 Prompting

Discrete prompting (DP). Following Schick and Schütze (2021), Le Scao and Rush (2021), we reformulate the NLI examples (cf. example in §1) into cloze-style questions using a human-designed

prompt. Specifically, we ask the PLM to fill in the blank (__) in sentence:

Premise . Question: Hypothesis ? Answer: __ .

Premise and Hypothesis are a pair of sentences from the NLI dataset. The gold labels are mapped to words in the PLM vocabulary. Concretely, we use following mapping (**verbalizer**; Schick and Schütze (2021)): “entailment”→ “yes”; “contradiction”→ “no”; “neutral”→ “maybe”. The optimization objective is to minimize the cross-entropy loss between the predicted and the gold words representing the three classes.

Soft prompting (SP; Li and Liang (2021); Qin and Eisner (2021); Zhong et al. (2021); Liu et al. (2021)) leverages prompts containing “pseudo tokens” that are not part of the PLM vocabulary. In this work, we ask a PLM to fill in the blank (__) in sentence:

Premise . Hypothesis ? $\langle v_1 \rangle \langle v_2 \rangle \langle v_3 \rangle \langle v_4 \rangle$ __ .

where each $\langle v_i \rangle$, $i \in \{1, 2, 3, 4\}$ is associated with a randomly initialized trainable vector (in the PLM’s lowest embedding layer) $v_i \in \mathbb{R}^d$, where d is the hidden dimension size of the embedding layer. Directly using v_i yields sub-optimal task performance: Li and Liang (2021) reparameterize v_i with another trainable matrix and then feed it forward through an MLP. Here, we adopt Liu et al. (2021)’s approach. They feed $[v_1, v_2, v_3, v_4]$ through an LSTM (Hochreiter and Schmidhuber, 1997) and use the outputs. PLM parameters, LSTM parameters, and v_i are jointly trained. Our SP and DP have the same training objective and verbalizer.

Mixed prompting (MP). We also experiment with a simple combination of DP and SP, by asking the PLM to fill in the blank (__) in sentence:

Premise . Question: Hypothesis ? $\langle v_1 \rangle \langle v_2 \rangle \langle v_3 \rangle \langle v_4 \rangle$
Answer: __ .

MP includes human descriptions of NLI as in DP and learns “soft prompts” as in SP.

3.3 Non-English prompting

We also explore the results of prompting the PLM with languages other than English, of which the *few-shot* annotation cost is affordable (Garrette and Baldrige, 2013; Lauscher et al., 2020; Zhao et al., 2021).

For a non-English language \mathcal{L} like Turkish, we translate¹ the English prompting words (§3.2)

¹We use Google Translate due to the simplicity of our

| | Prompt | Verbalizer |
|----|---|---------------------|
| EN | DP <u>Premise</u> . Question: <u>Hypothesis</u> ? Answer: _ . | Entailment → yes |
| | SP <u>Premise</u> . <u>Hypothesis</u> ? $\langle v_1 \rangle \dots \langle v_4 \rangle$ _ . | Contradict → no |
| | MP <u>Premise</u> . Question: <u>Hypothesis</u> ? $\langle v_1 \rangle \dots \langle v_4 \rangle$ Answer: _ . | Neutral → maybe |
| TR | DP <u>Premise</u> . Soru: <u>Hypothesis</u> ? Cevap: _ . | Entailment → Evet |
| | SP <u>Premise</u> . <u>Hypothesis</u> ? $\langle v_1 \rangle \dots \langle v_4 \rangle$ _ . | Contradict → hiçbir |
| | MP <u>Premise</u> . Soru: <u>Hypothesis</u> ? $\langle v_1 \rangle \dots \langle v_4 \rangle$ Cevap: _ . | Neutral → belki |

Table 1: Prompts and verbalizers in English (EN) and Turkish (TR). “ $\langle v_1 \rangle \dots \langle v_4 \rangle$ ” = “ $\langle v_1 \rangle \langle v_2 \rangle \langle v_3 \rangle \langle v_4 \rangle$ ”. Appendix §B shows translated prompts/verbalizers of the languages used in our experiments.

“Question” and “Answer” into \mathcal{L} , e.g., “Soru” and “Cevap” in Turkish. Correspondingly, the verbalizer maps gold labels into \mathcal{L} : “entailment” → “Evet”; “contradiction” → “hiçbir”; “neutral” → “belki”. Table 1 presents example prompts and verbalizers.

4 Dataset and Setup

Dataset. We conduct our experiments on natural language inference datasets MNLI and XNLI (Williams et al., 2018; Conneau et al., 2018). MNLI provides multi-genre English NLI sentence pairs. XNLI provides development and test splits of *human-translated parallel* NLI sentence pairs in 15 languages² and the *machine-translated* MNLI training sets in 14 languages.

For constructing the *few-shot training set*, we randomly sample without replacement $K \in \{1, 2, 4, 8, 16, 32, 64, 128, 256\}$ shots *per class* from the EN MNLI training split. Then we retrieve translations of this EN training set from XNLI to create the few-shot training sets in the other languages.

To simulate a realistic low-resource regime (Kann et al., 2019; Perez et al., 2021), we use *few-shot development sets*. For EN, we sample the same number of shots (as training) from the XNLI development split. As a result, a 2-shot experiment uses 2 training and 2 development shots per class. For other languages, we retrieve the translations of the English development set from XNLI. Following Conneau et al. (2018), we report *accuracy* on XNLI test.

Setup. We conduct all experiments using the pretrained XLM-RoBERTa-base model (Conneau et al., 2020) containing 270M parameters trained on 2.5 TB CommonCrawl data in 100 languages. We use PyTorch (Paszke et al., 2019) and the Hug-

prompt. Specialized bilingual dictionaries can also be used.

²The languages are English (EN), French (FR), Spanish (ES), German (DE), Greek (EL), Bulgarian (BG), Russian (RU), Turkish (TR), Arabic (AR), Vietnamese (VI), Thai (TH), Chinese (ZH), Hindi (HI), Swahili (SW), and Urdu (UR).

gingFace framework (Wolf et al., 2020).³

We use batch size 32 for finetuning and 24 for prompting methods due to resource limitations. Following Le Scao and Rush (2021), we use learning rate $1e-5$ for both finetuning and prompting. Following the suggestions of Mosbach et al. (2021), Zhang et al. (2021), we train the model with a large number of epochs (50) and select the checkpoint that performs best on the development set. We repeat each experiment 5 times with different random seeds ($\{1, 2, 3, 4, 5\}$) and report mean and variance. Appendix §A shows our reproducibility checklist.

5 Experiments

5.1 Zero-shot crosslingual transfer

We first compare prompting with finetuning in *zero-shot crosslingual transfer* (Pires et al., 2019; Conneau et al., 2020; Artetxe and Schwenk, 2019; Hu et al., 2020): The PLM is trained on the EN few-shot dataset and then directly evaluated on the test set of all languages. Table 2 reports the results.

EN results. From column *EN* we observe that: (i) As expected, all four methods benefit from more shots. (ii) Prompting methods (DP/SP/MP) clearly outperform finetuning especially in low-resource regimes. For example, in the 4-shot experiment, SP outperforms finetuning by ≈ 8 (41.84-33.90) accuracy points. Table 3 displays some examples for which SP outperforms finetuning. The improvements become less significant when more shots are available, e.g., 256. (iii) SP outperforms DP for most choices of shots (except 128), evidencing the strength of relaxing the “discrete token” constraint in DP (Liu et al., 2021; Qin and Eisner, 2021; Zhong et al., 2021). But we give up the interpretability of DP for this better performance. (iv) Performance of MP – the combination of DP and SP – is decent, but not stellar. Future work may explore advanced prompting methods succeeding

³Resources are available at <https://github.com/mprompting/xlmrprompt>

| Shots | Method | AR | BG | DE | EL | EN | ES | FR | HI | RU | SW | TH | TR | UR | VI | ZH | \bar{X} |
|-------|--------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| - | MAJ | 33.33 | 33.33 | 33.33 | 33.33 | 33.33 | 33.33 | 33.33 | 33.33 | 33.33 | 33.33 | 33.33 | 33.33 | 33.33 | 33.33 | 33.33 | 33.33 |
| 1 | FT | 32.53 | 32.63 | 32.94 | 32.53 | 32.91 | 32.61 | 32.65 | 32.87 | 32.67 | 32.77 | 33.11 | 32.68 | 32.87 | 32.69 | 32.77 | 32.75 |
| | DP | 32.08 | 33.23 | 32.97 | 33.24 | 33.15 | 33.78 | 34.08 | 33.41 | 33.78 | 33.45 | 33.00 | 34.01 | 31.99 | 32.83 | 33.64 | 33.24 |
| | SP | 34.84 | 36.50 | 36.87 | 37.49 | 36.65 | 38.29 | 38.57 | 36.43 | 37.56 | 34.52 | 35.71 | 34.76 | 35.54 | 35.06 | 37.61 | 36.43 |
| | MP | 32.31 | 32.32 | 33.03 | 32.14 | 33.29 | 34.02 | 33.74 | 34.12 | 33.03 | 32.86 | 32.18 | 34.59 | 32.65 | 32.82 | 33.35 | 33.10 |
| 2 | FT | 33.16 | 33.35 | 33.82 | 33.24 | 33.43 | 33.31 | 33.30 | 33.24 | 33.29 | 33.19 | 33.40 | 33.04 | 33.20 | 33.03 | 33.29 | 33.29 |
| | DP | 32.90 | 35.11 | 34.44 | 34.69 | 35.41 | 35.43 | 34.77 | 34.11 | 34.93 | 32.97 | 35.43 | 35.19 | 32.75 | 33.28 | 36.46 | 34.52 |
| | SP | 35.91 | 38.08 | 38.15 | 38.42 | 37.97 | 38.23 | 38.62 | 36.32 | 39.22 | 34.35 | 37.20 | 34.75 | 35.52 | 36.67 | 37.71 | 37.14 |
| | MP | 32.76 | 34.25 | 34.10 | 33.26 | 34.59 | 33.81 | 34.33 | 33.75 | 34.01 | 33.88 | 34.55 | 34.51 | 32.59 | 33.83 | 35.39 | 33.97 |
| 4 | FT | 33.86 | 33.89 | 33.73 | 33.63 | 33.90 | 33.58 | 33.55 | 33.86 | 33.58 | 33.75 | 33.71 | 33.79 | 33.67 | 33.85 | 33.78 | 33.74 |
| | DP | 35.42 | 37.64 | 38.85 | 37.67 | 39.50 | 38.91 | 38.26 | 36.43 | 37.54 | 34.72 | 37.76 | 37.23 | 35.92 | 36.02 | 38.74 | 37.37 |
| | SP | 38.04 | 40.46 | 40.08 | 40.79 | 41.84 | 39.78 | 41.10 | 37.55 | 41.72 | 35.81 | 39.23 | 35.88 | 37.66 | 37.86 | 39.48 | 39.15 |
| | MP | 33.14 | 33.79 | 35.16 | 33.95 | 36.26 | 35.52 | 35.44 | 34.63 | 34.21 | 33.53 | 35.96 | 35.62 | 33.51 | 34.06 | 37.10 | 34.79 |
| 8 | FT | 32.85 | 32.75 | 33.05 | 32.59 | 33.06 | 32.58 | 32.80 | 32.89 | 32.88 | 32.75 | 33.14 | 32.69 | 33.05 | 32.83 | 32.65 | 32.84 |
| | DP | 32.73 | 34.78 | 34.79 | 34.82 | 36.39 | 34.97 | 35.17 | 33.00 | 34.59 | 32.91 | 35.14 | 34.13 | 33.14 | 33.66 | 35.56 | 34.39 |
| | SP | 36.30 | 38.84 | 38.22 | 38.68 | 39.02 | 38.16 | 38.82 | 35.86 | 39.73 | 34.50 | 37.90 | 35.11 | 35.61 | 37.41 | 37.17 | 37.42 |
| | MP | 32.67 | 33.24 | 34.81 | 33.18 | 34.78 | 34.66 | 34.77 | 34.76 | 33.81 | 33.07 | 34.46 | 35.12 | 32.69 | 33.57 | 36.34 | 34.13 |
| 16 | FT | 33.72 | 34.09 | 34.28 | 33.49 | 34.73 | 33.82 | 33.81 | 33.08 | 34.06 | 33.69 | 33.06 | 33.57 | 33.22 | 34.01 | 33.46 | 33.74 |
| | DP | 35.07 | 37.07 | 37.51 | 37.43 | 38.24 | 36.91 | 36.61 | 35.85 | 36.51 | 33.84 | 37.21 | 35.74 | 34.86 | 35.77 | 37.86 | 36.43 |
| | SP | 38.88 | 40.60 | 40.21 | 40.44 | 39.45 | 39.37 | 40.90 | 36.86 | 40.61 | 37.11 | 39.45 | 36.26 | 35.88 | 38.46 | 37.35 | 38.79 |
| | MP | 32.46 | 33.02 | 33.98 | 32.59 | 33.20 | 34.54 | 34.39 | 34.30 | 33.90 | 33.28 | 33.47 | 34.69 | 32.67 | 33.28 | 35.68 | 33.70 |
| 32 | FT | 35.84 | 36.28 | 36.00 | 36.11 | 36.64 | 36.02 | 36.47 | 35.41 | 35.68 | 35.33 | 35.71 | 35.90 | 34.81 | 36.10 | 36.20 | 35.90 |
| | DP | 41.80 | 43.51 | 43.49 | 42.50 | 43.65 | 42.83 | 43.90 | 39.30 | 42.39 | 37.51 | 40.51 | 42.01 | 39.77 | 41.91 | 39.94 | 41.67 |
| | SP | 40.30 | 43.38 | 42.08 | 42.27 | 44.72 | 42.32 | 42.34 | 38.91 | 43.76 | 37.54 | 39.97 | 38.79 | 38.83 | 42.09 | 39.56 | 41.12 |
| | MP | 40.95 | 42.16 | 42.61 | 42.31 | 45.52 | 41.22 | 44.67 | 40.17 | 42.18 | 36.52 | 40.16 | 41.21 | 40.48 | 41.74 | 40.89 | 41.52 |
| 64 | FT | 40.16 | 39.56 | 40.10 | 39.87 | 41.68 | 40.34 | 39.47 | 39.53 | 38.34 | 39.64 | 39.18 | 39.50 | 39.23 | 40.85 | 39.63 | 39.81 |
| | DP | 45.64 | 47.64 | 48.05 | 46.94 | 48.89 | 44.95 | 47.97 | 41.61 | 44.85 | 40.98 | 45.65 | 45.67 | 43.37 | 47.30 | 45.24 | 45.65 |
| | SP | 43.48 | 43.81 | 45.99 | 43.70 | 49.04 | 45.79 | 46.11 | 40.86 | 44.51 | 40.49 | 44.68 | 41.91 | 40.09 | 45.25 | 44.17 | 43.99 |
| | MP | 43.86 | 46.01 | 48.22 | 46.79 | 51.84 | 46.61 | 48.31 | 40.11 | 44.75 | 37.84 | 45.01 | 44.82 | 43.95 | 48.28 | 43.03 | 45.30 |
| 128 | FT | 43.50 | 45.52 | 45.60 | 44.38 | 46.94 | 45.75 | 46.00 | 42.96 | 44.94 | 41.43 | 43.27 | 43.67 | 41.78 | 44.81 | 44.79 | 44.36 |
| | DP | 46.23 | 50.49 | 50.99 | 47.39 | 53.68 | 48.53 | 49.28 | 44.77 | 46.93 | 42.03 | 47.95 | 49.56 | 44.21 | 48.92 | 49.56 | 48.03 |
| | SP | 44.78 | 46.24 | 45.30 | 46.31 | 49.45 | 45.80 | 46.37 | 43.29 | 44.95 | 41.21 | 45.64 | 41.93 | 41.18 | 44.99 | 45.73 | 44.88 |
| | MP | 46.48 | 47.98 | 49.04 | 49.09 | 52.55 | 49.66 | 50.34 | 47.03 | 46.40 | 42.89 | 48.08 | 48.45 | 44.04 | 48.15 | 50.47 | 48.04 |
| 256 | FT | 52.13 | 54.57 | 54.43 | 54.00 | 57.79 | 55.89 | 55.39 | 50.65 | 52.90 | 50.00 | 51.22 | 52.31 | 48.57 | 54.16 | 52.10 | 53.07 |
| | DP | 53.23 | 55.59 | 55.39 | 55.05 | 60.14 | 50.64 | 54.43 | 46.10 | 51.35 | 45.26 | 53.42 | 50.83 | 48.42 | 55.14 | 52.72 | 52.51 |
| | SP | 52.26 | 56.04 | 53.02 | 53.12 | 60.58 | 54.80 | 55.79 | 49.43 | 52.49 | 47.33 | 54.52 | 52.08 | 48.48 | 54.54 | 54.59 | 53.27 |
| | MP | 52.77 | 53.98 | 50.71 | 54.63 | 60.13 | 51.64 | 55.32 | 49.58 | 53.50 | 45.27 | 53.37 | 51.28 | 47.16 | 52.34 | 53.80 | 52.37 |

Table 2: Zero-shot crosslingual transfer results in accuracy (%). Each number is the mean performance of 5 runs, when using finetuning (FT), discrete prompting (DP), soft prompting (SP), and mixed prompting (MP). “MAJ”: majority baseline; \bar{X} : macro average across 15 languages. Please see Appendix Table 7 for variances.

| Premise/Hypothesis | Prediction |
|---|--------------------|
| This was the temper of the times. This wasn't the temper of the times. | “no” (Contradict) |
| We would go in there. We would enter there at 8pm. | “maybe” (Neutral) |
| I hope to hear from you soon. I hope we talk soon. | “yes” (Entailment) |

Table 3: Qualitative examples for which prompting outperforms finetuning.

in both task performance and interpretability. We focus on DP and SP in following experiments.

Crosslingual transfer results closely follow the trends of EN results: Prompting outperforms finetuning when looking at the macro average \bar{X} . One intriguing finding is that DP successfully transfers the learned knowledge to target languages, better than SP in some languages, using the *code-switched* prompt: “Premise . Question: Hypothesis ? Answer: _ . ” where Premise and Hypothesis are non-English. Thus, DP is able to leverage the strong crosslingual ability of the

multilingual PLM. Like finetuning, prompting does not uniformly benefit the 14 non-English languages. For example, the crosslingual transfer performance of HI/SW/UR is notably inferior compared with other languages.

Overall, prompting outperforms finetuning in zero-shot crosslingual transfer of NLI in the low-resource regimes.

5.2 In-language prompting

We next compare prompting with finetuning when using non-English few-shot datasets. Taking Turkish as an example, recall that we can use the Turkish prompts (§3.3) and few-shot datasets from XNLI (§4) to finetune/prompt the PLM directly.

Table 4 shows results of in-language experiments of Turkish, Urdu, Swahili, and Chinese. We make two main observations: (i) Prompting still outperforms finetuning, though the non-English prompts and verbalizers are translated from EN simply using Google Translate. (ii) In-language results are slightly worse but competitive to transfer learning results (Table 2). We conjecture that

| Shots | Method | TR | UR | SW | ZH |
|-------|--------|--------------|--------------|--------------|--------------|
| 8 | FT | 32.71 | 32.83 | 32.80 | 33.31 |
| | DP | 38.02 | 39.33 | 33.84 | 37.46 |
| | SP | 35.41 | 34.59 | 33.47 | 34.39 |
| 16 | FT | 33.00 | 33.78 | 33.46 | 33.56 |
| | DP | 39.39 | 40.58 | 34.48 | 42.24 |
| | SP | 40.22 | 35.47 | 33.99 | 35.64 |
| 32 | FT | 37.15 | 34.23 | 34.52 | 35.38 |
| | DP | 48.79 | 41.67 | 37.52 | 38.04 |
| | SP | 43.62 | 39.18 | 36.00 | 35.13 |
| 64 | FT | 38.87 | 35.90 | 36.37 | 42.14 |
| | DP | 48.97 | 42.34 | 37.72 | 44.73 |
| | SP | 47.26 | 39.12 | 37.98 | 40.75 |
| 128 | FT | 40.84 | 36.23 | 36.81 | 43.16 |
| | DP | 49.73 | 45.22 | 41.26 | 49.24 |
| | SP | 47.68 | 40.96 | 40.42 | 47.17 |
| 256 | FT | 49.41 | 40.12 | 42.17 | 48.98 |
| | DP | 52.61 | 46.10 | 47.69 | 53.11 |
| | SP | 51.21 | 44.60 | 46.89 | 52.76 |

Table 4: In-language results in accuracy (%). Prompting (DP/SP) outperforms finetuning (FT). Please see Appendix Table 6 for variances.

two factors result in the second observation. First, some languages have a small amount of pretraining data. For example, Swahili has 1.6GB pretraining data while English has 300GB (Conneau et al., 2020). Thus, the PLM may not be well pretrained for solving tasks in Swahili directly. Second, the few-shot training data for non-English languages is machine-translated (§4). With better *few-shot* translations and in-language expertise, prompting possibly could achieve even better results.

Overall, the experimental results show that directly prompting PLMs with non-English languages is also an effective way of solving NLU tasks in low-resource regimes.

6 Conclusion

We showed that prompting performs better than finetuning in few-shot crosslingual transfer and in-language training of multilingual natural language inference. We hope our results will encourage more research about prompting multilingual tasks and models.

Future work may explore using text-to-text models like T5 (Raffel et al., 2020) or other autoregressive PLMs (Lewis et al., 2020). Investigating PLMs containing even more parameters is also promising as shown by Lester et al. (2021).

Acknowledgements

We thank the anonymous reviewers for the insightful comments and suggestions. This work was funded by the European Research Council (ERC #740516).

References

- Mikel Artetxe and Holger Schwenk. 2019. [Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond](#). *Transactions of the Association for Computational Linguistics*, 7:597–610.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. [A large annotated corpus for learning natural language inference](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. [Xnli: Evaluating cross-lingual sentence representations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. [The pascal recognising textual entailment challenge](#). In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

- Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.
- Dan Garrette and Jason Baldridge. 2013. [Learning a part-of-speech tagger from two hours of annotation](#). In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 138–147, Atlanta, Georgia. Association for Computational Linguistics.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. [Long short-term memory](#). *Neural Comput.*, 9(8):1735–1780.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. [XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421, Virtual. PMLR.
- Katharina Kann, Kyunghyun Cho, and Samuel R. Bowman. 2019. [Towards realistic practices in low-resource natural language processing: The development set](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3342–3349, Hong Kong, China. Association for Computational Linguistics.
- Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. [From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.
- Teven Le Scao and Alexander Rush. 2021. [How many data points is a prompt worth?](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online. Association for Computational Linguistics.
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Xiang Lisa Li and Percy Liang. 2021. [Prefix-tuning: Optimizing continuous prompts for generation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.
- Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. GPT understands, too. *arXiv preprint arXiv:2103.10385*.
- Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.
- Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. [On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines](#). In *International Conference on Learning Representations*.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. [Pytorch: An imperative style, high-performance deep learning library](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *arXiv preprint arXiv:2105.11447*.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. [Language models as knowledge bases?](#) In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. [How multilingual is multilingual BERT?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.
- Guanghui Qin and Jason Eisner. 2021. [Learning how to ask: Querying LMs with mixtures of soft prompts](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*,

- pages 5203–5212, Online. Association for Computational Linguistics.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Ori Ram, Yuval Kirstain, Jonathan Berant, Amir Globerson, and Omer Levy. 2021. [Few-shot question answering by pretraining span selection](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3066–3079, Online. Association for Computational Linguistics.
- Timo Schick and Hinrich Schütze. 2021. [It’s not just size that matters: Small language models are also few-shot learners](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.
- Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. [AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.
- Derek Tam, Rakesh R Menon, Mohit Bansal, Shashank Srivastava, and Colin Raffel. 2021. Improving and simplifying pattern exploiting training. *arXiv preprint arXiv:2103.11955*.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. [Superglue: A stickier benchmark for general-purpose language understanding systems](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. [Xlnet: Generalized autoregressive pretraining for language understanding](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Wenpeng Yin, Nazneen Fatema Rajani, Dragomir Radev, Richard Socher, and Caiming Xiong. 2020. [Universal natural language processing with limited annotations: Try few-shot textual entailment as a start](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8229–8239, Online. Association for Computational Linguistics.
- Mo Yu, Xiaoxiao Guo, Jinfeng Yi, Shiyu Chang, Saloni Potdar, Yu Cheng, Gerald Tesauro, Haoyu Wang, and Bowen Zhou. 2018. [Diverse few-shot text classification with multiple metrics](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1206–1215, New Orleans, Louisiana. Association for Computational Linguistics.
- Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2021. [Revisiting few-sample BERT fine-tuning](#). In *International Conference on Learning Representations*.
- Mengjie Zhao, Tao Lin, Fei Mi, Martin Jaggi, and Hinrich Schütze. 2020. [Masking as an efficient alternative to finetuning for pretrained language models](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2226–2241, Online. Association for Computational Linguistics.
- Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Ivan Vulić, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2021. [A closer look at few-shot crosslingual transfer: The choice of shots matters](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5751–5767, Online. Association for Computational Linguistics.
- Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. [Factual probing is \[MASK\]: Learning vs. learning to recall](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, Online. Association for Computational Linguistics.

A Reproducibility Checklist

A.1 Model architecture and number of parameters

We use the `xlm-roberta-base` model (Conneau et al., 2020). It contains 12 Transformer blocks with 768 hidden dimensions. Each block has 12 attention heads. Vocabulary size is 250K. Overall the model has 270M parameters and was pretrained on the on 2.5 TB of newly created clean CommonCrawl data in 100 languages.

Following Liu et al. (2021), we employ a bidirectional LSTM in SP and MP. The hidden dimension is also 768 so the number of LSTM parameters is $2 \times 4 \times (768 \times 768 + 768 \times 768 + 768) \approx 10M$. Another MLP is used to project the concatenation of LSTM states back to 768-dimension which has $(768 \times 2 \times 768 + 768) \approx 1.2M$. SP and MP also have four learnable vectors v_i resulting in $768 \times 4 = 3072$ parameters.

A.2 Computing infrastructure

All experiments are conducted on GeForce GTX 1080Ti. For finetuning, we use batch size 32 and 4 GPUs. Because prompting uses the masked language model objective so we use a maximum batch size 24. A single GPU is used for 1-shot experiments. Two and three GPUs are used for 2- and 4-shot experiments. Other experiments use 6 GPUs.

A.3 Evaluation metrics

Our code is available at <https://github.com/mprompting/xlmrprompt>. We use the standard evaluation metric accuracy as Conneau et al. (2018). For finetuning, evaluation script path is `./finetuning/utils/eval_meters.py`. For DP, evaluation script path is `./pet/pet/trainers/meters.py`. For SP/MP, evaluation script path is `./sptuning/pet/trainers/meters.py`.

A.4 Hyperparameter search

We use the same learning rate ($1e-5$) as Le Scao and Rush (2021) who compare prompting and finetuning in English NLU tasks. No learning rate scheduling is used for clear comparisons. For both finetuning and prompting, the model is trained for 50 epochs and the checkpoint that performs best on development set is selected for performance evaluation.

| | Prompt | Verbalizer |
|----|--|---------------------------------|
| TR | DP <code>P . Soru: H ? Cevap: _ .</code> | Entailment \rightarrow Evet |
| | SP <code>P . H ? <v1>...<v4> _ .</code> | Contradict \rightarrow hiçbir |
| | MP <code>P . Soru: H ? <v1>...<v4> Cevap: _ .</code> | Neutral \rightarrow belki |
| SW | DP <code>P . Swali: H ? Jibu: _ .</code> | Entailment \rightarrow ndio |
| | SP <code>P . H ? <v1>...<v4> _ .</code> | Contradict \rightarrow hasi |
| | MP <code>P . Swali: H ? <v1>...<v4> Jibu: _ .</code> | Neutral \rightarrow labda |
| ZH | DP <code>P . 问题: H ? 答案: _ .</code> | Entailment \rightarrow 是 |
| | SP <code>P . H ? <v1>...<v4> _ .</code> | Contradict \rightarrow 否 |
| | MP <code>P . 问题: H ? <v1>...<v4> 答案: _ .</code> | Neutral \rightarrow 也许 |

Table 5: Prompts and verbalizers in Turkish (TR), Swahili (SW), and Chinese (ZH).

| Shots | Method | TR | UR | SW | ZH |
|-------|--------|------------------|------------------|------------------|------------------|
| 8 | FT | 32.71 \pm 0.61 | 32.83 \pm 0.29 | 32.80 \pm 0.56 | 33.31 \pm 0.27 |
| | DP | 38.02 \pm 1.14 | 39.33 \pm 0.58 | 33.84 \pm 0.44 | 37.46 \pm 0.62 |
| | SP | 35.41 \pm 0.30 | 34.59 \pm 0.26 | 33.47 \pm 0.34 | 34.39 \pm 0.25 |
| 16 | FT | 33.00 \pm 0.93 | 33.78 \pm 0.58 | 33.46 \pm 0.91 | 33.56 \pm 0.50 |
| | DP | 39.39 \pm 0.81 | 40.58 \pm 0.67 | 34.48 \pm 0.86 | 42.24 \pm 2.66 |
| | SP | 40.22 \pm 0.50 | 35.47 \pm 0.61 | 33.99 \pm 0.17 | 35.64 \pm 1.03 |
| 32 | FT | 37.15 \pm 1.78 | 34.23 \pm 0.85 | 34.52 \pm 1.20 | 35.38 \pm 0.47 |
| | DP | 48.79 \pm 0.40 | 41.67 \pm 1.39 | 37.52 \pm 1.08 | 38.04 \pm 1.00 |
| | SP | 43.62 \pm 0.67 | 39.18 \pm 1.09 | 36.00 \pm 1.23 | 35.13 \pm 0.75 |
| 64 | FT | 38.87 \pm 0.99 | 35.90 \pm 1.26 | 36.37 \pm 1.13 | 42.14 \pm 1.22 |
| | DP | 48.97 \pm 0.56 | 42.34 \pm 0.91 | 37.72 \pm 0.81 | 44.73 \pm 0.86 |
| | SP | 47.26 \pm 0.77 | 39.12 \pm 1.36 | 37.98 \pm 1.48 | 40.75 \pm 1.90 |
| 128 | FT | 40.84 \pm 1.45 | 36.23 \pm 0.19 | 36.81 \pm 1.85 | 43.16 \pm 0.95 |
| | DP | 49.73 \pm 0.73 | 45.22 \pm 0.57 | 41.26 \pm 1.48 | 49.24 \pm 0.89 |
| | SP | 47.68 \pm 0.68 | 40.96 \pm 1.23 | 40.42 \pm 1.53 | 47.17 \pm 0.54 |
| 256 | FT | 49.41 \pm 2.03 | 40.12 \pm 1.77 | 42.17 \pm 1.77 | 48.98 \pm 3.12 |
| | DP | 52.61 \pm 1.34 | 46.10 \pm 0.40 | 47.69 \pm 1.12 | 53.11 \pm 0.61 |
| | SP | 51.21 \pm 0.30 | 44.60 \pm 0.91 | 46.89 \pm 0.81 | 52.76 \pm 0.29 |

Table 6: In-language results in accuracy (%). Prompting (DP/SP) outperforms finetuning (FT). We report mean and variance of 5 runs.

A.5 Datasets and preprocessing

We retrieve the MNLI and XNLI datasets from the official websites: cims.nyu.edu/~sbowman/multinli and cims.nyu.edu/~sbowman/xnli. We use the tokenizer in the HuggingFace framework (Wolf et al., 2020) to preprocess the texts. In all experiments, the max sequence length is 256.

B Translated Prompts

Table 5 shows the prompts and verbalizers used in in-language experiments. We use Google Translate but more specialized bilingual dictionaries can also be used. For Urdu, we show the prompt and verbalizer in the code repository.

C More Results

Table 6 and Table 7 show performances with variances.

| Shots | Method | AR | BG | DE | EL | EN | ES | FR | HI | RU | SW | TH | TR | UR | VI | ZH | \bar{X} |
|-------|--------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| - | MAJ | 33.33 | 33.33 | 33.33 | 33.33 | 33.33 | 33.33 | 33.33 | 33.33 | 33.33 | 33.33 | 33.33 | 33.33 | 33.33 | 33.33 | 33.33 | 33.33 |
| 1 | FT | 32.53±1.08 | 32.63±1.02 | 32.94±0.50 | 32.53±1.10 | 32.91±0.62 | 32.61±0.89 | 32.65±0.95 | 32.87±0.64 | 32.67±0.94 | 32.77±0.67 | 33.11±0.26 | 32.68±0.88 | 32.87±0.59 | 32.69±0.88 | 32.77±0.86 | 32.75±0.16 |
| | DP | 32.08±0.21 | 32.23±0.15 | 32.97±0.25 | 33.24±0.10 | 33.15±0.16 | 33.78±0.08 | 34.08±0.05 | 33.41±0.07 | 33.78±0.10 | 33.45±0.36 | 33.00±0.30 | 34.01±0.07 | 31.99±0.27 | 32.83±0.27 | 33.64±0.06 | 33.24±0.60 |
| | SP | 34.84±1.67 | 36.50±1.41 | 36.87±0.48 | 37.49±0.37 | 36.65±1.52 | 38.29±0.40 | 38.57±0.34 | 38.57±0.34 | 37.56±1.30 | 34.52±0.61 | 35.71±1.78 | 34.76±0.34 | 35.54±0.28 | 35.06±1.65 | 37.61±0.40 | 36.43±1.27 |
| | MP | 32.31±0.20 | 32.32±0.15 | 33.03±0.15 | 32.14±0.27 | 33.29±0.60 | 34.02±0.11 | 33.74±0.34 | 34.12±0.22 | 33.03±0.42 | 32.86±0.25 | 32.18±0.17 | 34.59±0.27 | 32.65±0.16 | 32.82±0.21 | 33.35±0.06 | 33.10±0.73 |
| 2 | FT | 33.16±0.56 | 33.35±1.13 | 33.82±0.70 | 33.24±0.82 | 33.43±1.12 | 33.31±0.91 | 33.30±1.18 | 33.24±0.69 | 33.29±1.12 | 33.19±0.84 | 33.40±0.57 | 33.04±0.87 | 33.20±0.53 | 33.03±0.62 | 33.29±0.60 | 33.29±0.18 |
| | DP | 32.90±0.74 | 35.11±0.53 | 34.44±0.38 | 34.69±0.51 | 35.41±0.55 | 35.43±0.65 | 34.77±0.55 | 34.11±0.60 | 34.93±0.51 | 32.97±0.37 | 35.43±0.35 | 35.19±0.59 | 32.75±1.22 | 33.28±0.94 | 36.46±0.65 | 34.52±1.07 |
| | SP | 35.91±2.09 | 38.08±1.41 | 38.15±0.85 | 38.42±0.99 | 37.97±1.42 | 38.23±0.53 | 38.62±0.52 | 36.32±0.47 | 39.22±1.43 | 34.35±0.44 | 37.20±1.56 | 34.75±0.40 | 35.52±0.66 | 36.67±1.64 | 37.71±0.74 | 37.14±1.43 |
| | MP | 32.76±0.21 | 34.25±0.40 | 34.10±0.69 | 33.26±0.55 | 34.59±0.50 | 33.81±0.37 | 34.33±0.44 | 33.75±0.23 | 34.01±0.38 | 33.88±0.36 | 34.55±0.78 | 34.51±0.10 | 32.59±0.60 | 33.83±0.34 | 35.39±1.06 | 33.97±0.69 |
| 4 | FT | 33.86±1.15 | 33.89±1.34 | 33.73±1.12 | 33.63±0.67 | 33.90±1.22 | 33.58±0.94 | 33.55±1.05 | 33.86±1.48 | 33.58±1.27 | 33.75±1.11 | 33.71±1.28 | 33.79±1.43 | 33.67±0.98 | 33.85±1.53 | 33.78±1.54 | 33.74±0.12 |
| | DP | 35.42±0.46 | 37.64±0.39 | 38.85±0.50 | 37.67±0.53 | 39.50±0.37 | 38.91±0.44 | 38.26±0.23 | 36.43±0.38 | 37.54±0.30 | 34.72±0.24 | 37.76±0.56 | 37.23±0.33 | 35.92±0.53 | 36.02±0.56 | 38.74±0.40 | 37.37±1.86 |
| | SP | 38.04±1.81 | 40.46±1.67 | 40.08±1.23 | 40.79±1.42 | 41.84±2.10 | 39.78±1.35 | 41.10±2.03 | 37.55±1.05 | 41.72±1.99 | 35.81±1.49 | 39.23±1.27 | 35.88±1.30 | 37.66±1.49 | 37.86±1.56 | 39.48±1.55 | 39.15±1.88 |
| | MP | 33.14±0.49 | 33.79±1.26 | 35.16±1.93 | 33.95±1.17 | 36.26±2.38 | 35.52±0.95 | 35.44±1.23 | 34.63±0.73 | 34.21±1.10 | 33.53±0.74 | 35.96±1.36 | 35.62±1.19 | 33.51±0.73 | 34.06±0.89 | 37.10±0.77 | 34.79±1.13 |
| 8 | FT | 32.85±0.44 | 32.75±0.42 | 33.05±0.46 | 32.59±0.54 | 33.06±0.82 | 32.58±0.39 | 32.80±0.38 | 32.89±0.34 | 32.88±0.37 | 32.75±0.24 | 33.14±0.38 | 32.69±0.66 | 33.05±0.27 | 32.83±0.41 | 32.65±0.31 | 32.84±0.17 |
| | DP | 32.73±0.81 | 34.78±0.63 | 34.79±0.56 | 34.82±0.66 | 36.39±0.43 | 34.97±0.90 | 35.17±0.29 | 33.00±0.50 | 34.59±0.43 | 32.91±0.25 | 35.14±0.60 | 34.13±0.60 | 33.14±0.54 | 33.66±0.82 | 35.56±0.51 | 34.39±1.05 |
| | SP | 36.30±0.94 | 38.84±0.68 | 38.22±0.34 | 38.68±0.69 | 39.02±0.57 | 38.16±0.86 | 38.82±1.08 | 35.86±0.53 | 39.73±0.45 | 34.50±0.69 | 37.90±0.71 | 35.11±1.05 | 35.61±1.22 | 37.41±0.98 | 37.17±0.64 | 37.42±1.54 |
| | MP | 32.67±0.70 | 33.24±0.67 | 34.81±1.45 | 33.18±1.13 | 34.78±2.22 | 34.66±1.05 | 34.77±0.60 | 34.76±0.94 | 33.81±0.37 | 33.07±0.21 | 34.46±0.82 | 35.12±0.41 | 32.69±0.76 | 33.57±0.50 | 36.34±0.82 | 34.13±1.01 |
| 16 | FT | 33.72±0.84 | 34.09±0.74 | 34.28±0.66 | 33.49±1.15 | 34.73±0.97 | 33.82±0.88 | 33.81±1.13 | 33.08±0.38 | 34.06±1.23 | 33.69±0.88 | 33.06±0.78 | 33.57±0.93 | 33.22±0.45 | 34.01±0.63 | 33.46±0.91 | 33.74±0.44 |
| | DP | 35.07±1.55 | 37.07±2.07 | 37.51±1.91 | 37.43±2.39 | 38.24±2.15 | 36.91±1.67 | 35.85±1.26 | 35.85±1.67 | 36.51±1.89 | 33.84±0.78 | 37.21±1.45 | 35.74±1.12 | 34.86±1.75 | 35.77±1.94 | 37.86±1.44 | 36.43±1.18 |
| | SP | 38.88±0.92 | 40.60±0.43 | 40.21±0.66 | 40.44±0.87 | 39.45±0.71 | 39.37±0.55 | 40.90±0.43 | 36.86±0.59 | 40.61±0.53 | 37.11±0.61 | 39.45±0.43 | 36.26±1.04 | 35.88±0.47 | 38.46±0.70 | 37.35±0.95 | 38.79±1.65 |
| | MP | 32.46±0.19 | 33.02±0.43 | 33.98±1.01 | 32.59±0.34 | 33.20±0.62 | 34.54±0.59 | 34.39±0.62 | 34.30±0.19 | 33.90±0.51 | 33.28±0.56 | 33.47±1.57 | 34.69±0.40 | 32.67±0.42 | 33.28±0.34 | 35.68±1.87 | 33.70±0.88 |
| 32 | FT | 35.84±1.36 | 36.28±1.47 | 36.00±0.95 | 36.11±1.32 | 36.64±1.46 | 36.02±1.49 | 36.47±1.27 | 35.47±1.26 | 35.68±1.03 | 35.33±1.27 | 35.71±0.98 | 35.90±1.31 | 34.81±1.24 | 36.10±1.59 | 36.20±1.62 | 35.90±0.45 |
| | DP | 41.80±0.85 | 43.51±0.94 | 43.49±0.72 | 42.50±0.37 | 43.65±0.46 | 42.83±0.66 | 43.90±0.79 | 39.30±1.68 | 42.39±0.83 | 37.51±0.75 | 40.51±1.02 | 42.01±0.76 | 39.77±1.18 | 41.91±0.60 | 39.94±1.57 | 41.67±1.81 |
| | SP | 40.30±1.73 | 43.38±0.52 | 42.08±0.49 | 42.27±1.00 | 44.72±0.90 | 42.32±0.78 | 42.34±0.77 | 38.91±1.53 | 43.76±0.74 | 37.54±1.18 | 39.97±1.20 | 38.79±0.75 | 38.83±0.45 | 42.09±1.27 | 39.56±1.12 | 41.12±2.06 |
| | MP | 40.95±1.18 | 42.16±0.97 | 42.61±1.03 | 42.31±0.70 | 45.52±0.56 | 41.22±0.62 | 44.67±1.23 | 40.17±0.68 | 42.18±0.85 | 36.52±0.92 | 40.16±1.17 | 41.21±0.95 | 40.48±0.51 | 41.74±0.83 | 40.89±0.95 | 41.52±1.99 |
| 64 | FT | 40.16±2.05 | 39.56±1.83 | 40.10±2.42 | 39.87±1.67 | 41.68±2.26 | 40.34±2.62 | 39.47±2.01 | 39.53±2.11 | 38.34±1.68 | 39.64±1.60 | 39.18±1.81 | 39.50±1.91 | 39.23±1.87 | 40.85±2.39 | 39.63±2.12 | 39.81±0.75 |
| | DP | 45.64±0.59 | 47.64±0.81 | 48.05±0.66 | 46.94±0.51 | 48.89±0.95 | 44.95±0.85 | 47.97±0.68 | 41.61±0.77 | 44.85±0.61 | 40.98±1.14 | 45.65±0.84 | 45.67±1.15 | 43.37±0.49 | 47.30±0.72 | 45.24±1.17 | 45.65±2.23 |
| | SP | 43.48±0.71 | 43.81±0.92 | 45.99±0.56 | 43.70±0.22 | 49.04±0.46 | 45.79±0.91 | 46.11±0.86 | 40.86±0.81 | 44.51±1.26 | 40.49±0.35 | 44.68±0.34 | 41.91±1.50 | 40.09±0.64 | 45.25±0.72 | 44.17±0.86 | 43.99±2.33 |
| | MP | 43.86±1.05 | 46.01±1.13 | 48.22±1.20 | 46.79±1.39 | 51.84±1.12 | 46.61±1.39 | 48.31±0.86 | 40.11±1.23 | 44.75±1.24 | 37.84±0.53 | 45.01±1.02 | 44.82±1.22 | 43.95±1.25 | 48.28±1.16 | 43.03±1.50 | 45.30±3.33 |
| 128 | FT | 43.50±1.77 | 45.52±2.29 | 45.60±2.40 | 44.38±2.56 | 46.94±2.98 | 45.75±2.75 | 46.00±1.91 | 42.96±2.64 | 44.94±2.28 | 41.43±1.88 | 43.27±2.02 | 43.67±2.56 | 41.78±2.46 | 44.81±2.44 | 44.79±1.67 | 44.36±1.52 |
| | DP | 46.23±1.06 | 50.49±0.84 | 50.99±0.75 | 47.39±0.80 | 53.68±0.48 | 48.53±1.00 | 49.28±0.56 | 44.77±0.95 | 46.93±0.54 | 42.03±0.68 | 47.95±0.66 | 49.56±0.71 | 44.21±0.84 | 48.92±1.36 | 49.56±0.77 | 48.03±2.83 |
| | SP | 44.78±4.43 | 46.24±3.94 | 45.30±4.21 | 46.31±3.70 | 49.45±5.63 | 45.80±4.61 | 46.37±4.48 | 43.29±4.19 | 44.95±3.74 | 41.21±3.54 | 45.64±3.95 | 41.93±4.95 | 41.18±4.09 | 44.99±4.54 | 45.73±4.93 | 44.88±2.13 |
| | MP | 46.48±2.40 | 47.98±1.98 | 49.04±1.99 | 49.09±1.32 | 52.55±0.86 | 49.66±1.52 | 50.34±1.28 | 47.03±1.40 | 46.40±2.25 | 42.89±1.95 | 48.08±1.65 | 48.45±1.25 | 44.04±1.99 | 48.15±2.19 | 50.47±1.55 | 48.04±2.38 |
| 256 | FT | 52.13±2.07 | 54.57±1.95 | 54.43±2.26 | 54.00±1.63 | 57.79±2.36 | 55.89±2.37 | 55.39±1.97 | 50.65±1.71 | 52.90±2.03 | 50.00±1.89 | 51.22±1.92 | 52.31±2.06 | 48.57±2.25 | 54.16±1.71 | 52.10±2.60 | 53.07±2.35 |
| | DP | 53.23±0.56 | 55.59±1.40 | 55.39±1.43 | 55.05±0.86 | 60.14±0.67 | 50.64±0.83 | 54.43±1.29 | 46.10±1.77 | 51.35±0.86 | 45.26±1.07 | 53.42±0.86 | 50.83±0.71 | 48.42±1.10 | 55.14±0.72 | 52.72±0.68 | 52.51±3.76 |
| | SP | 52.26±1.89 | 56.04±1.71 | 53.02±2.14 | 53.12±1.62 | 60.58±0.86 | 54.80±1.42 | 55.79±1.15 | 49.43±1.87 | 52.49±1.78 | 47.33±1.11 | 54.52±1.21 | 52.08±1.49 | 48.48±0.93 | 54.54±1.05 | 54.59±1.31 | 53.27±3.17 |
| | MP | 52.77±0.31 | 53.98±1.26 | 50.71±2.41 | 54.63±0.95 | 60.13±0.40 | 51.64±2.83 | 55.32±0.79 | 49.58±1.87 | 53.50±0.68 | 45.27±0.70 | 53.37±0.23 | 51.28±1.26 | 47.16±0.75 | 52.34±1.15 | 53.80±0.82 | 52.37±3.38 |

Table 7: Zero-shot crosslingual transfer results in accuracy (%). We report mean and variance of 5 runs, when using finetuning (FT), discrete prompting (DP), soft prompting (SP), and mixed prompting (MP). “MAJ”: majority baseline; \bar{X} : macro average across 15 languages.